Fake News Detection and Manipulation Reasoning via Large Vision-Language Models

Ruihan Jin
Department of Automation, Tsinghua University
China
jrh20@mails.tsinghua.edu.cn

Ruibo Fu†
Institute of Automation, Chinese Academy of Sciences
China
ruibo.fu@nlpr.ia.ac.cn

Zhengqi Wen
Institute of Automation, Chinese Academy of Sciences
China
zqwen@nlpr.ia.ac.cn

Shuai Zhang
Department of Automation, Tsinghua University
China
zhang_shuai@mail.tsinghua.edu.cn

Yukun Liu
School of Artificial Intelligence, University of Chinese Academy of Sciences
China
yukunliu927@gmail.com

Jianhua Tao
Department of Automation, Tsinghua University
China
jhtao@tsinghua.edu.cn

ABSTRACT

Fake news becomes a growing threat to information security and public opinion with the rapid sprawl of media manipulation. Therefore, fake news detection attracts widespread attention from academic community. Traditional fake news detection models demonstrate remarkable performance on authenticity binary classification but their ability to reason detailed faked traces based on the news content remains under-explored. Furthermore, due to the lack of external knowledge, the performance of existing methods on fact-related news is questionable, leaving their practical implementation unclear. In this paper, we propose a new multi-media research topic, namely manipulation reasoning. Manipulation reasoning aims to reason manipulations based on news content. To support the research, we introduce a benchmark for fake news detection and manipulation reasoning, referred to as Human-centric and Fact-related Fake News (HFFN). The benchmark highlights the centrality of human and the high factual relevance, with detailed manual annotations. HFFN encompasses four realistic domains with fake news samples generated through three manipulation approaches. Moreover, a Multi-modal news Detection and Reasoning langUage Model (M-DRUM) is presented not only to judge on the authenticity of multi-modal news, but also raise analytical reasoning about potential manipulations. On the feature extraction level, a cross-attention mechanism is employed to extract fine-grained fusion features from multi-modal inputs. On the reasoning level, a large vision-language model (LVLM) serves as the backbone to facilitate fact-related reasoning. A two-stage training framework is deployed to better activate the capacity of identification and reasoning. Comprehensive experiments demonstrate that our model outperforms state-of-the-art (SOTA) fake news detection models and powerful LVLMs like GPT-4 and LLaVA.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; *Social networks.*

KEYWORDS

Large vision-language model, Fake news detection, Benchmark, Multi-modal learning

ACM Reference Format:
Ruihan Jin, Ruibo Fu†, Zhengqi Wen, Shuai Zhang, Yukun Liu, and Jianhua Tao. 2018. Fake News Detection and Manipulation Reasoning via Large Vision-Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (MM'24)Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The development of online media greatly improves the convenience of information communication. On the contrast, recent years witness a rampancy of disinformation, which poses threat to information security and public opinion. News is at enormous risk of manipulation for being a common carrier of multi-modal information, which draws attention within the academia community and various fake news detection methods are proposed. Early works on fake news detection prioritize the identification of uni-modal manipulation. Currently, with the advent of deep generative models, media manipulation expands across multiple modalities. Visual deepfake models can edit human faces and generate high-fidelity images and videos [24, 37]. With large language models (LLMs) like BERT [8] and GPT [23], lexical replacement and editing is performed easily to modify semantics and facts. Media manipulation enhances the difficulty of detection and has a more detrimental social impact when it targets human-centric and fact-related news.

† Corresponding author.

MM'24, October 28 - November 1, 2024, Melbourne, Australia.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

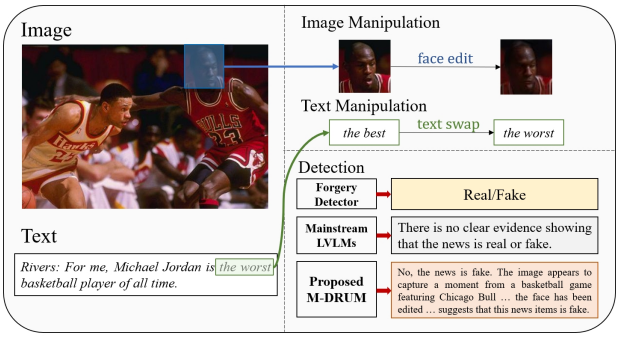


Figure 1: An illustration of multi-modal fake new detection and manipulation reasoning. We construct Human-centric and Fact-related Fake News(HFFN) benchmark through three approaches of media manipulation. We proposed Multi-modal news Detection and Reasoning langUage Model(M-DRUM) to not only perform authenticity classification but also reason about manipulations.

To address multi-modal media manipulation, modern fake news detection approaches leverage feature-level interaction of different modalities [3, 32, 39]. Despite the favourable performance achieved, two major challenges still exist. First, most forgery detectors fail to reason about potential manipulations. Mere authenticity binary classification is trivial for analyzing manipulations or sorting out forgery mechanism, which limits the practical implementation. Second, as mentioned above, manipulations tend to attack human-centric news involving celebrities or well-known events, with a high factual relevance. It is critical to identify human-centric and fact-related fake news to eliminate negative social impacts.

In this paper,a new multi-media research task is proposed, namely manipulation reasoning. Manipulation reasoning aims to reason about potential manipulations based on news content. Existing benchmarks failed to provide analytical reasoning about the manipulation on news and lack the bias toward human-centric and fact-related news. To facilitate further research, we present a benchmark for fake news detection and manipulation reasoning, which is designed for both forgery detectors and general-purposed LVLMs. The benchmark is referred to as **H**uman-centric and **F**act-related **F**ake **N**ews(**HFFN**). Specifically, HFFN collects multi-modal news represented by image-text pairs, encompassing four domains: *entertainment*, *sport*, *politics* and *others*. News samples in HFFN emphasize the centrality of human and high factual relevance. Three manipulation approaches are developed to perform multi-modal fake news generation. Furthermore, detailed manual annotations are attached to news to facilitate manipulation reasoning.

As illustrated in Fig.1, neither existing fake news detection models nor mainstream LVLMs achieve satisfactory results on multi-modal news. Performing fake news detection and manipulation reasoning urges a combination of authenticity representation and general knowledge. Owning a wealth of general knowledge, large vision-language models (LVLMs) are cut out for manipulation reasoning. In this paper, we proposed **M**ulti-modal news **D**etection and **R**easoning lang**U**age **M**odel(**M-DRUM**), a novel large vision-language model for multi-modal fake news detection. M-DRUM aligns images and texts with a multi-modal encoder and leverages a manipulation-specific facial feature to enhance human-centric representation. Multi-modal features are aggregated through cross-modal fusion and are prompted to an LVLM to generate detection results and analytical reasoning. To our best knowledge, we are the first to employ LVLM as the backbone model for fake news detection. Comprehensive experimental results demonstrate that M-DRUM outperforms SOTA multi-modal fake news detection models and mainstream LVLMs. The enhancement of few-shot learning and chain-of-thought reasoning is confirmed with further experiments.

Main contributions of this paper are as follows:

- We present the fake news detection and manipulation reasoning benchmark HFFN. The benchmark is constructed following the principle of "*human-centric*" and "*fact-related*". Fake news samples in HFFN are generated through three manipulation approaches and encompass four realistic domains.
- We propose M-DRUM, a novel large vision-language model for fake news detection and manipulation reasoning. M-DRUM not only detect authenticity classes based on the multi-modal news, but also perform analytical reasoning about potential manipulations.
- Comprehensive experiments demonstrate M-DRUM outperforms SOTA multi-modal fake news detection models and powerful LVLMs like GPT-4 and LLaVA both quantitatively and qualitatively. Further experiments verify the improvement of few-shot learning and chain-of-thought reasoning.

2 RELATED WORK

2.1 Media Manipulation.

Disinformation becomes a growing threat to information security and public opinion with the rampancy of media manipulation. Media manipulation methods varies across different modalities. In visual modality, GAN-based methods are widely employed to manipulate human faces with text-guidance [19, 21] or latent space editing [30, 36]. [24] utilizes the multi-modal semantics to guide the editing process. [37] enables high-fidelity image inversion and attribute editing by a distortion consultation approach. In textual modality, common manipulation methods include conditional text generation [2, 28] and text style transfer [33, 35]. Recent progress in natural language generation gives rise to large-scale manipulable text [6]. Manipulations toward human-centric and fact-related news may cause harmful impact to society. In our work, by applying off-the-shelf manipulation methods, we build a multi-modal fake news benchmark following the principle of "*human-centric*" and "*fact-related*" to evaluate detection and reasoning.

2.2 Fake News Detection.

Fake news detection draws great attention as news is at enormous risk of multi-modal manipulation. Social context based detection methods judge on the authenticity of news based on the spreading procedure such as social network [22] and post-user interaction [20]. Content-based methods differentiate fake and real news by finding manipulation cues [25, 26]. Recent researches focus on identifying

multi-modal news. [39] proposes an effective textual and visual feature fusion method with co-attention. [3, 40] leverage the adaptable aggregation between uni-modal and cross-modal features to resolve the inherent ambiguity across different modalities. [13] introduces LLM as a data augmentation approach to generate advisable rationales for subsequent detection. Different from aforementioned methods, we propose a novel architecture combining feature extraction and LVLM for fake news detection and manipulation reasoning. To our best knowledge, we are the first to employ LVLM as the backbone model for fake news detection.

2.3 Large Vision-Language Models.

Expanding the multi-modal capability of LLMs is a current research focus. [16] employs Flan-T5 [5] with a Q-Former to bridge the modality gap between visual feature and language model. [34] leverages the combination of ImageBind [10] and Vicuna [4] to deal with multi-modal input. By instruction tuning on multi-modal instruction-following data generated by GPT-4 [1], [17, 18] achieve impressive cross-modal chat abilities. Despite the general knowledge derived from large-scale pre-training, these models lack domain-specific expertise. To better prompt LVLMs with manipulation detection expertise, we introduce a multi-level prompt learner to enhance manipulation reasoning. Fig.1 exhibits M-DRUM outperforms existing forgery detectors and LVLMs with profound manipulation detection expertise and broad general knowledge.

3 HFFN: HUMAN-CENTRIC AND FACT-RELATED FAKE NEWS BENCHMARK

3.1 Design Principles

Human-Centric Human-centric news carries a higher risk of manipulation than other news topics. High-fidelity deepfake models can perform face swap and facial attribute editing easily, posing harmful threats to visual authenticity. Human centric news is highlighted in the construction of our benchmark, which means we pay higher attention to news samples with clear human faces. Images with no faces or blurred faces are filtered out. To simulate potential image manipulations, face swap and facial-attribute editing are both conducted to create fake images.

Fact-Related Factual errors are common in media manipulation, which result in misleading public opinion and negative social impact. Traditional fake news detection methods have difficulty distinguishing factual errors due to the lack of general knowledge or external knowledge source. There is a strong demand to measure whether a detection model is capable of tackling factual factual manipulation. Our benchmark is tailored for LVLMs and contains sufficient fact-related news samples. During data collection, we gather news featuring celebrities and well-known events as we believe these news is at a higher risk of being factually manipulated. After collection, random factual errors are added to the news to examine the capacity of the detection model.

3.2 Construction Process: Data Collection

The original news samples are gathered from latest real-world media sources. Among them, human-centric and fact-related samples get the most attention. Following the design principles, news without clear human faces or high factual relevance is screened out. The screening process is carefully conducted by multiple volunteers. To enhance validity, we calculate the image-text consistency of news by CLIP [27]. News with low image-text consistency are removed to improve the validity of the benchmark. To this end, the original news samples set $O = \{I_{real}, T_{real}\}$ is obtained.

3.3 Construction Process: Media Manipulation

Image Manipulation Inspired by manipulation procedure of DGM[4], we achieve image manipulation with face swap and face editing. Face swap manipulation refers to replace the main character's face with another person's. InfoSwap [9] is adopted to swap faces by replacing the largest face f_o in the original image I_o with a random source face f_{swap} from CelebA-HQ dataset [14]. Face editing manipulation refers to modify the facial attributes of the main character. For example, we intentionally put a smiling face or render an exaggerated beard on his/her face. We achieve high-fidelity editing effects with a GAN-based method [37] to transfer the original face f_o into target style f_{edit}. In both ways, the manipulated face is stitched back to the original image I_o and the manipulated image I_s is get. The bounding box of the manipulated region $b = \{x_1, y_1, x_2, y_2\}$ is recorded as annotation.

Text Manipulation In text manipulation, the textual semantics are attacked by word substitution. Assisted by ChatGPT [23], words in the input headline T_o are revised to reverse the global semantics of the text. For example, an original headline of "Liu Xiang returns triumphantly and receives heated extolling" is altered as "Liu Xiang returns triumphantly and receives harsh questioning". Therefore, the global semantic of the headline is reversed.

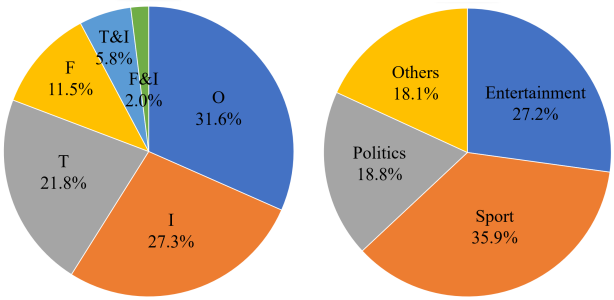
Factual Manipulation Among various factual manipulation methods for text, entity replacement is one of the most common and convenient, which is adopted to create samples with factual errors. Specifically, given the input headline T_o, a Named Entity Recognition(NER) model [7] is launched to extract the name of the main entity. The extracted name is replaced with a randomly chosen name subsequently. We record the manipulated text derived from text manipulation and factual manipulation as T_s.

Three uni-modal manipulation approaches mentioned above are conducted on the original dataset O by randomly alter the original samples I_o, T_o with manipulated ones I_s, T_s. A total of five manipulation types are formed, including three uni-modal types(Image, Text and Fact) and two cross-modal types(Image&Text and Image&Fact).

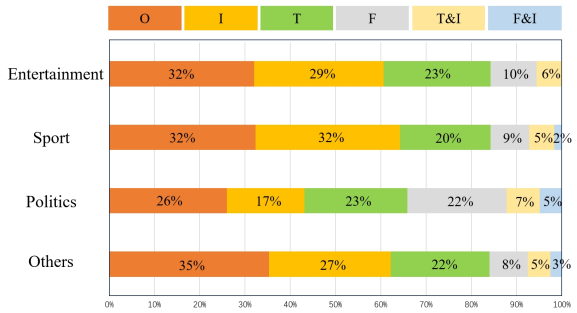
3.4 Construction Process: Human Annotation

In HFFN benchmark, we annotate the multimodal news samples with detailed evidence for manipulation reasoning. Annotating HFFN is a challenging task, which requires annotators to endow with background knowledge in the relevant field and provide analytical reasoning based on details of news. We hire 10 professional annotators to annotate the reasoning process with following steps:

1) *Authenticity annotation.* Point out the authenticity of the in terms of "Yes, the news is real" or "No, the news is fake", given the manipulation type of the news.



(a) Manipulation types in HFFN. (b) Four domains in HFFN.



(c) Distribution of different manipulation types

Figure 2: Statistics of HFFN benchmark. (I: Image Manipulation, T: Text Manipulation, F: Factual Manipulation, &: combination of two manipulation types)

2) *Content summary.* Raise a description of news contents for specific content analysis. The description includes the perspective of image content, headline description and image-text consistency.

3) *Clue revelation.* Provide detailed clues or reasoning for unveiling manipulations such as traces of face editing or factual errors in the headline.

During annotation, we provide news contents and corresponding manipulation types for annotators and they are required to annotate the reasoning process. Each news item is shown to two annotators to perform independent annotation. Comparing the quality of their annotations, the final annotation is selected from them.

3.5 Overview of HFFN

The overall statistics of HFFN are visualized in Fig.2. In line with the design principles, HFFN benchmark attaches great significance to the human-centric and fact-related news. HFFN consists a total of 655 samples, encompassing four realistic domains: *entertainment*, *sport*, *politics* and *others*. Each news sample is represented as an image-text pair, equipped with detailed manual annotation. The average length of manual annotations is 69.1 tokens. The overall manipulation rate of HFFN is 68.4%, including 7.8% of multi-modal manipulation samples.

4 M-DRUM: MULTI-MODAL NEWS DETECTION AND REASONING LANGUAGE MODEL

To address fake news detection and manipulation reasoning, as illustrated in Fig.3, we present M-DRUM, a novel large vision-language model based architecture. In M-DRUM, we use a multi-modal encoder to extract visual and textual features from news images and headlines. We leverage a cross-attention mechanism to obtain multi-modal fusion features. A prompt learner bridges the gap between manipulation expertise and the general knowledge of LVLM and based on that, a LVLM generates the analytical reasoning. The model is trained under a two-stage framework to strengthen the capacity of identification and reasoning.

4.1 Multi-modal Feature Extraction

Driven by the idea of multi-modal alignment, we use ImageBind [10], a powerful cross-modal alignment model as the feature encoder. Given the news image $I \in \mathbb{R}^{H\times W\times C}$ and the corresponding headline T, we firstly extract visual and textual features with ImageBind. Inspired by AnomalyGPT [11], we obtained 4 intermediate visual features $F^i_{image} \in \mathbb{R}^{H_i\times W_i\times C_{image}}$ from encoding at different depths, where i indicates the i-th depth. Accordingly, the textual feature $F_{text} \in \mathbb{R}^{C_{text}}$ is extracted from the headline.

To thoroughly comprehend multi-modal inputs, a cross-modal fusion is adopted to integrate uni-modal features. Focusing on the cross-modal relationship, the cross-modal features $F^i_{cross} \in \mathbb{R}^{C_{fusion}}$ are obtained by calculating the cross-attention score between the rectified visual feature and the textual feature in the softmax-free linear attention [15] expressions. The cross-modal fusion process can be represented as:

$$\begin{aligned} \tilde{F}^i_{image} &= \text{LinearLayer}(F^i_{image}), \\ F^i_{cross} &= \tilde{F}^i_{image} \cdot F^T_{text}, \end{aligned} \tag{1}$$

where i indicates the i-th stage.

Face encoder In our work, human-centric news is highlighted which suffers from face swap and malicious editing. Therefore, identification aids provided by facial authenticity features are necessary. In M-DRUM, we leverage a face encoder to extract the manipulation-specific representation of human faces. The face encoder is modeled with a ResNet-50 [12] and pretrained on large-scale Deepfake dataset [31] to provide feature-level guidance for identifying facial authenticity. The extracted facial feature F_{face} is then concatenated with the cross-modal features to obtain the ultimate fusion feature. The fusion process can be represented as:

$$F_{fusion} = \text{Concate}(F^i_{cross}, F_{face}). \tag{2}$$

4.2 Manipulation Reasoning

In M-DRUM, a LVLM serves as a mighty knowledge base for reasoning generation. To prompt LVLM with the authenticity of the news and inspire the general knowledge, we design a hybrid prompt learner to bridge the gap between the manipulation expertise and the general knowledge of LVLM. The prompt learner aims to assist the LVLM in understanding the manipulation information of multi-modal news comprehensively. As shown in Fig.3, the prompt learner integrates three parts of information. The fusion feature

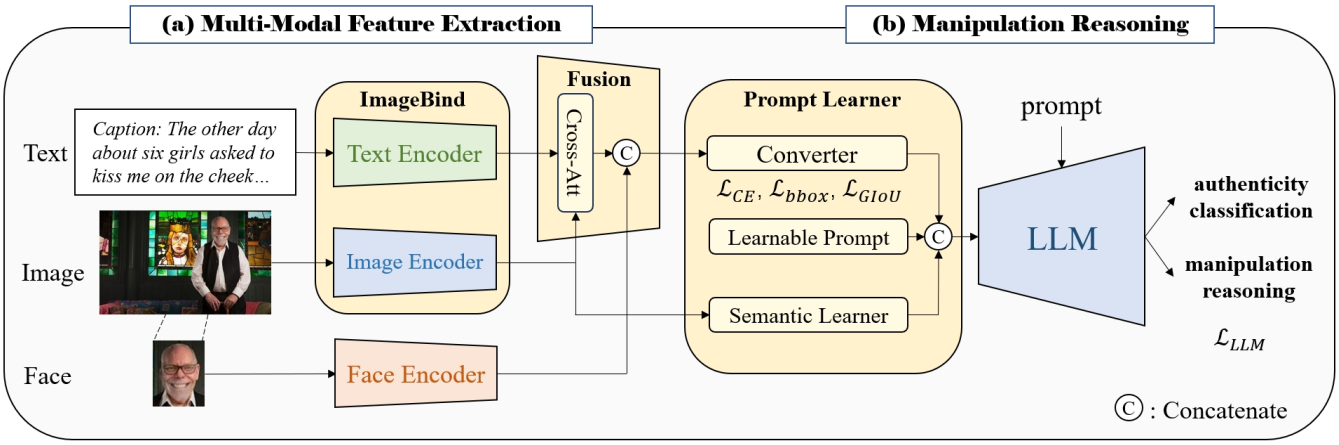


Figure 3: The architecture of M-DRUM. In M-DRUM, news images and headlines are aligned with a multi-modal encoder and a manipulation-specific facial feature is leveraged to enhance human-centric representation. Fusion features are derived with the cross-attention mechanism. To bridge the gap between the manipulation expertise and the general knowledge of LVLM, a prompt learner is adopted and a LVLM raises authenticity classification and manipulation reasoning. The model is trained under a two-stage framework to strengthen the capacity of identification and reasoning.

F_{fusion} is transformed into prompt embeddings with a converter. A prediction head is introduced to provide specific guidance to the conversion process, supervised by the authenticity label of the news, together with bounding box of the edited regions. We expect the LVLM to accept semantic information from news images as much as possible so that it can be combined with facts for reasoning. To better transform the semantics information, we leverage a semantic learner to derive visual semantics from the visual features F_{image} in the form of prompt embeddings. Additionally, to learn specific prompts for manipulation reasoning in a self-adaptive way, learnable prompt embeddings are adopted. The multi-level prompts are fed to the LVLM to raise analytical reasoning about potential manipulations. The global function of the prompt learner is:

$$E_{prompt} = \text{Concate}(C(F_{fusion}|H(F_{fusion})), L(F_{image}), E_{ada}), \tag{3}$$

where C, H, L stand for the converter, the prediction head and the semantic learner respectively. E_{ada} stands for the self-adaptive prompt embeddings.

4.3 Loss Functions for Performance Augmentation

To augment fake news detection and reasoning, we constrain our model with three types of loss functions: cross-entropy loss, bounding box loss and GIoU loss.

Cross-entropy Loss Cross-entropy loss is widely used in classification and natural language generation tasks. For authenticity classification, cross-entropy loss is introduced to supervise the prediction head in the prompt learner, which is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}}[y_i \log(p_i) + (1-y_i)\log(1-p_i)], \tag{4}$$

where $\mathcal{B}$ is the batch size, y_i is the authenticity label of the i-th sample and p_i is the probability for positive prediction.

For manipulation reasoning, cross-entropy loss quantifies the disparity between the generated reasoning and the target text sequence, which is defined as:

$$\mathcal{L}_{LLM} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(p_i), \tag{5}$$

where n is the number of tokens, t_i is the ground truth label for token i and q_i is the predicted probability for token i.

Bounding box Loss We utilize L1 loss to supervise manipulated regions predicted by the prediction head, which is defined as:

$$\mathcal{L}_{bbox} = \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} |b_{pred} - b_{gt}|, \tag{6}$$

where $\mathcal{B}$ is the batch size, b_{pred} and b_{gt} is the predicted and true bounding boxes respectively.

GIoU Loss Intersection over Union(IoU) loss is commonly used in object detection tasks with scale invariance. GIoU [29] serves as an improvement of IoU by optimizing in the case of non-overlapping bounding boxes, which is defined as:

$$\mathcal{L}_{GIoU} = 1 - \text{GIoU} = 1 - (\text{IoU} - \frac{|A^c - \mathcal{U}|}{A^c}), \tag{7}$$

where A^c and $\mathcal{U}$ is the smallest enclosing box and the union area of the predicted and the true bounding boxes respectively. We introduce bounding box loss and GIoU loss to assist our model in locating manipulated regions and understanding visual semantics.

The global loss function can be calculated by the weighting of each loss functions:

$$\mathcal{L} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{LLM} + \gamma\mathcal{L}_{bbox} + \delta\mathcal{L}_{GIoU}, \tag{8}$$

where $\alpha, \beta, \gamma, \delta$ are hyper-parameters.

4.4 Two-Stage Training Process

To better combine the capabilities of multi-modal feature extraction and manipulation reasoning, we adopt a two-stage framework to train M-DRUM, refers to detection learning and reasoning learning.

Detection Learning In the detection learning stage, we set the face encoder and the prompt learner to be trainable. We train our model on the large-scale multi-modal media manipulation dataset DGM^4 [32]. During the training process, we expect the model to improve the performance of authenticity classification in large-scale detection task, which serves as the basis for subsequent manipulation reasoning.

Reasoning Learning In the reasoning learning stage, only the prompt learner is trainable. The training is launched on a delicately annotated human-centric and fact-related fake news detection benchmark HFFN. At this stage, we expect our model to improve analysis and reasoning abilities with multi-level prompts and generate analytical reasoning with confidence and vividness.

5 EXPERIMENTS

In this section, we evaluate the performance of M-DRUM on HFFN benchmark. Quantitative results on authenticity classification demonstrate M-DRUM outperforms SOTA multi-modal fake news detection models. The effectiveness of few-shot learning and chain-of-thought reasoning is also verified. Qualitative analysis on manipulation reasoning exhibits that M-DRUM proposes reasoning with more confidence and vividness compared with mainstream LVLMs.

5.1 Experimental Settings

Baselines Quantitative and qualitative experiments are performed on fake news detection and manipulation reasoning, respectively. For fake news detection, baselines are MCAN [39] and HAMMER [32], which are the state-of-the-art multi-modal detection models. For manipulation reasoning, we compare our method with powerful LVLMs including PandaGPT, GPT-4 and LLaVa by prompting them to propose reasoning on potential media manipulations.

Metrics In quantitative experiments, we evaluate models on accuracy, precision, recall and F1-score, which are commonly used in fake news detection tasks to measure the performance of authenticity classification. In qualitative experiments, we evaluate the reasoning results manually. 12 independent human raters are employed to assess the quality of randomly chosen reasoning results. Human evaluation is conducted on three orthogonal aspects: *Exactness*, *Certainty* and *Detail*. *Exactness* refers to whether the reasoning results are correct and consistent with the news content. *Certainty* refers to whether the reasoning results are clear and an ambiguous answer is regarded as a low score. *Detail* refers to whether the reasoning results are analyzed in detail rather than talk in generalities. Human raters are asked to score the reasoning on a scale of 1 to 10, with higher scores indicating better performance. We show the average evaluation scores on three aspects of M-DRUM and mainstream LVLMs.

Implementation Details We use ImageBind-Huge [10] as the multi-modal feature encoder and Vicuna-7B[4] as the LVLM. We concatenate the outputs from the 8th, 16th, 24th, 32nd layers of the image encoder into the visual feature. The parameters of the model are initialized using the pre-trained parameters in PandaGPT [34].

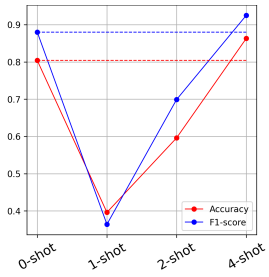


Figure 4: Performance floating of few-shot learning.

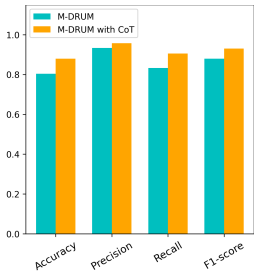


Figure 5: Efficacy of chain-of-thought (CoT) reasoning.

In both the detection and reasoning learning stage, training is performed with a learning rate of 1e-3 and a batch size of 16. We set the image resolution to be 224×224. Loss weights $\alpha, \beta, \gamma, \delta$ are set to 1 by default. All of our experiments are conducted on 2 NVIDIA 3090 GPUS with PyTorch framework.

5.2 Authenticity Classification

We compare M-DRUM with SOTA multi-modal fake news detection models MCAN and HAMMER on HFFN benchmark. The results are presented in Tab.1. By comparison, M-DRUM exceeds the performance of the baselines on HFFN benchmark. Specifically, M-DRUM achieves the highest accuracy of 80.4%, exceeding the SOTA fake news detection models by 8.2%. In four domains of HFFN, M-DRUM ranks either first or second in terms of precision, recall and F1-score. Quantitative results exhibits the superiority of M-DRUM to identify the human-centric and fact-related news. The performance advantage of M-DRUM on HFFN highly relies on the emphasis of facial feature and the general knowledge owned by the LVLM to which feature-based detection models are not comparable.

Few-shot Learning Considering the implementation of LVLM, we expect the M-DRUM to perform better under few-shot learning [2] settings. We test the performance of M-DRUM in 0, 1, 2 and 4-shot learning settings. The result is shown in Tab.3 and Fig.4. As the prompt examples adding, the classification performance of M-DRUM declines and then climb up. This can be explained with: a small number of prompt examples tend to confuse the model, and a larger number of samples promote the model to synthesize implicit rules and make correct judgements. Furthermore, 4-shot learning outperforms 0-zero by 5.9% in accuracy, indicating an appropriate number of examples leads to better performance, which verifies the promotion of few-shot learning.

Chain-of-Thought Reasoning A chain-of-thought (CoT) refers to a series of intermediate reasoning steps that mimic the reasoning process of human and significantly promote LLMs to tackle complex tasks [38]. In HFFN, manual annotations of *content summary* and *clue revelation* serves as the intermediate steps of the CoT. We explore whether the reasoning ability of M-DRUM can be improved through a step-by-step process in few-shot learning. In the CoT strategy, the *content summary* and the *clue revelation* of news are added to training examples and M-DRUM is guided to reason based on both original news content and manually annotated intermediate steps. Fig.5 demonstrates the performance of CoT reasoning, where M-DRUM receives a 7.6% accuracy boost and a 5.0% F1-score boost

Method	Acc.	Entertainment			Sport			Politics			Others		
		Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
MCAN	0.710	0.896	0.729	0.804	0.892	0.776	0.830	0.811	0.652	0.723	0.850	0.667	0.747
HAMMER	0.722	0.951	0.639	0.765	0.913	**0.913**	**0.913**	0.818	0.643	0.720	**0.914**	0.653	0.762
M-DRUM	**0.804**	**0.980**	**0.820**	**0.893**	**0.951**	0.853	0.899	**0.872**	**0.810**	**0.840**	0.913	**0.840**	**0.875**

Table 1: Comparison among multi-modal fake news detection models on HFFN. The best-performing model is in bold.

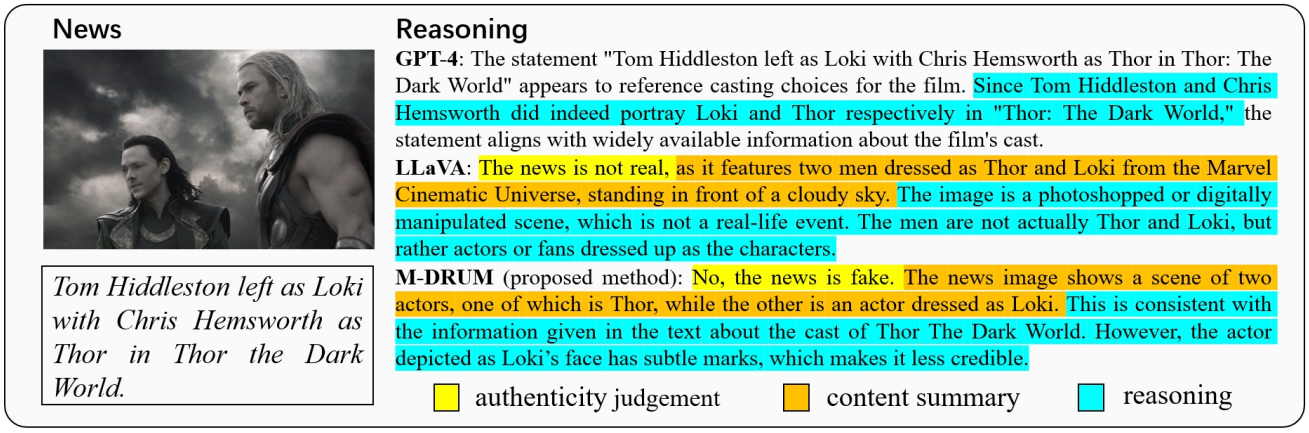


Figure 6: An example of the manipulation reasoning proposed by M-DRUM, compared with powerful LVLMs. The content analysis of the reasoning is marked with different colors.

Method	Human Evaluation			Total
	Exactness	Certainty	Detail	
GT	**9.30**	**8.97**	**8.48**	**8.92**
PandaGPT	1.42	4.65	1.82	2.63
GPT-4	2.33	2.80	4.60	3.24
LLaVA	2.03	7.17	**7.40**	5.53
M-DRUM	**9.10**	**8.45**	7.25	**8.27**

Table 2: Human evaluation on the manipulation reasoning proposed by M-DRUM and powerful LVLMs(scores range from 1~10). The best-performing model is in bold. (GT: human annotation)

Setup	Accuracy	Precision	Recall	F1-score
0-shot	0.804	**0.934**	0.833	0.880
1-shot	0.396	0.647	0.253	0.364
2-shot	0.596	0.810	0.615	0.699
4-shot	**0.863**	0.885	**0.968**	**0.925**

Table 3: Few-shot fake news detection results on HFFN. The best performance is highlighted in bold.

with CoT instruction. The result shows that generating the CoT along with the answer benefits the detection of M-DRUM.

5.3 Manipulation Reasoning

We encourage the detection model to propose analytical reasoning about manipulation in assist of unveiling forgery mechanisms. To evaluate the reasoning results, we scored M-DRUM and mainstream LVLMs on a manual basis. Tab.2 shows the result of human evaluation. In terms of *exactness* and *certainty*, M-DRUM far exceeds other LVLMs. Slightly inferior to LLaVA, reasoning results proposed by M-DRUM are still rich in *detail*. In general, the analytical reasoning proposed by M-DRUM can reach the performance of the ground truth (manual annotations). An example of the manipulation reasoning is shown in Fig.6. Compared with the detailed manipulation reasoning generated by M-DRUM, the reasoning of GPT-4 is ambiguous and the reasoning of LLaVA is multi-leveled but flawed. Conclusively, M-DRUM raises analytical reasoning with more confidence and vividness, which greatly expands the implementation of fake news detection models.

5.4 Ablation Studies

To evaluate the role of each modality in fake news detection and verify the effectiveness of architecture design, ablation experiments are conducted. We explore the impact of ignoring visual, textual and facial features in detection. In each set of experiment, a certain modality of M-DRUM is eliminated by removing the corresponding feature encoder and the two-stage training process is re-conducted. We collated the classification results of the ablation model which are presented in Tab.4. It can be observed that all ablation models with certain modality eliminated suffers from severe performance degradation, which proves the importance of each modality in

Method	Image	Text	Face	Accuracy	Precision	Recall	F1-score
M-DRUM	✓	✓	✓	**0.804**	**0.934**	**0.833**	**0.880**
M-DRUM w.o I		✓	✓	0.569	0.839	0.580	0.686
M-DRUM w.o T	✓		✓	0.388	0.571	0.475	0.519
M-DRUM w.o F	✓	✓		0.745	0.890	0.794	0.840

Table 4: Results of ablation studies on modalities of M-DRUM. The ✓indicates module inclusion.

M-DRUM and the necessity of facial authenticity features toward human-centric news detection.

6 CONCLUSION

In this paper, we introduce the fake news detection and manipulation reasoning benchmark HFFN. The benchmark is constructed following the principles of "*human-centric*" and "*fact-related*". To address classification and reasoning, we present M-DRUM, a novel detection model leveraging LVLM as the backbone. Combining multi-modal manipulation expertise and the general knowledge of LVLM, M-DRUM can not only perform authenticity judgement on the multi-modal news, but also enable analytical reasoning about potential manipulations. Comprehensive experiments demonstrate that M-DRUM outperforms SOTA fake news detection models and mainstream LVLMs. Further experiments verify the improvement of few-shot learning and chain-of-thought reasoning. Ablation studies exhibit the indispensability of different modals in detection.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown, Mann Benjamin, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[6] Evan Crothers, Nathalie Japkowicz, and Herna L Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection models. *IEEE Access* (2023).

[7] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. *arXiv preprint arXiv:2106.01760* (2021).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. 2021. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3404–3413.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[11] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2023. Anomalygpt: Detecting industrial anomalies using large vision-language models. *arXiv preprint arXiv:2308.15366* (2023).

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *arXiv preprint arXiv: 2309.12247* (2023).

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[15] Soroush Abbasi Koohpayegani and Hamed Pirsiavash. 2024. Sima: Simple softmax-free attention for vision transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2607–2617.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[19] Yahui Liu, Marco De Nai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. 2020. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1357–1365.

[20] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022*. 1148–1158.

[21] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems* 31 (2018).

[22] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1165–1174. https://doi.org/10.1145/3340531.3412046

[23] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI Blog* (2022). https://openai.com/blog/chatgpt

[24] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

[25] Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 490–497.

[26] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Sastry Girish, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 658–666.

[30] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2287–2296.

[31] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

[32] Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[33] Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*. 187–193.

[34] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355* (2023).

[35] Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. In *EMNLP-IJCNLP*.

[36] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.

[37] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Cenference on Computer Vision and Pattern Recognition*. 11379–11388.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[39] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2560–2569.

[40] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 343–352.

Supplementary Materials: Fake News Detection and Manipulation Reasoning via Large-Vision Language Models

Anonymous Authors

1 THE DATA FORMAT OF HFFN BENCHMARK

In our work, We propose the benchmark of **H**uman-centric and **F**act-related **F**ake **N**ews (**HFFN**). The data format of HFFN are illustrated in Tab.1 with descriptions. Encompassing four domains and five manipulation types, news with both image and text modalities was presented with detailed manual annotations. By facilitating the benchmark with detailed human annotations, we expect to leverage HFFN for evaluating the performance on authenticity classification and analytical reasoning of fake news detection models and general-purposed LVLMs.

2 IMPLEMENTATION OF THE BASELINE METHODS

As the supplements to *section 5.1*, all the baseline methods were implemented with their publicly available source codes on 2 NVIDIA 3090 GPUS with PyTorch framework. We compare our model with SOTA detection models including MCAN and HAMMER in fake news detection task. For MCAN, we pre-trained it on large-scale multi-modal media manipulation dataset DGM^4 and carefully tuned it on HFFN to achieve optimal performance. For HAMMER, we utilized the default settings provided in the original paper. In manipulation reasoning task, our model is compared with mainstream LVLMs including PandaGPT, GPT-4 and LLaVA. Specifically, the open-source parameters of `pandagpt_7b_max_len_512` and `llava-v1.5-7b` were used to launch PandaGPT and LLaVA respectively. We designed over 10 prompt templates for manipulation reasoning respectively and pre-test is conducted to evaluate which template leads to better performance on manipulation reasoning. Considering LVLMs are large-scale black-box models, we evaluate the potential of the templates from the quality of the inference results. The template for the best results is chosen as the final prompt template. We present the prompts used in our experiments to evaluate the manipulation reasoning ability of M-DRUM as follows:

Prompts for manipulation reasoning <Img> E_{img} </Img> Assume you are an expert in manipulation reasoning. This is a photo selected from a piece of news, which needs to be real and consistent with the headline of the news: {headline}. The news may be confronted with media manipulations. You are required to reason about manipulations of the news. The reasoning needs to be consistent to the news content, and to be clear and detailed. Reasoning result: {answer}

Prompts for PandaGPT, GPT-4 and LLaVA are similar to this with slightly differences. They are not listed here due to space constraints.

3 PERFORMANCE OF FEW-SHOT LEARNING

The detailed inference results of M-DRUM under few-shot learning settings are illustrated in Fig.1. Under the measurement of each indicators, the classification performance of M-DRUM declines and then climb up as the prompt examples adding.

Keys	Description
image	The image of news.
text	The headline of news.
domain	The domain of news. (one of *entertainment*, *sport*, *politics* and *others*)
label	The authenticity label of news. (0 stands for real and 1 stands for fake.)
fake_cls	The manipulation class of news. (*orig* for real news or one of the five manipulation types, including three uni-modal types and two cross-modal types.)
face_bbox	The bounding box of the main human face, which is regarded as the fake region if the news image is manipulated.
reasoning	Human annotated result of the analytical reasoning on manipulations.

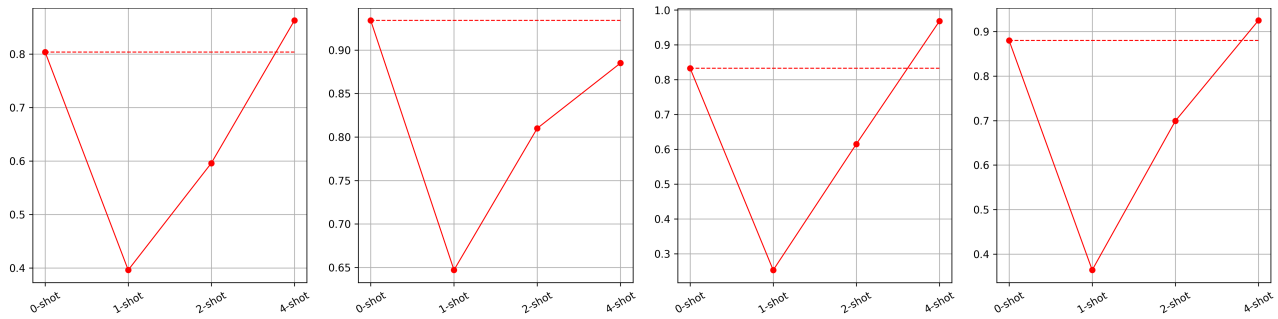
Table 1: Data formats of HFFN benchmark.

4 MORE EXAMPLES OF MANIPULATION REASONING

We compared our model with mainstream LVLMs in *section 5.3*, specifically selecting PandaGPT, GPT-4 and LLaVA on manipulation reasoning. More examples of the comparative analysis on HFFN are presented in Fig.2, Fig.3, Fig.4 and Fig.5. Outside the scope of HFFN benchmark, we conduct a small-scale test on the DGM^4 dataset with M-DRUM and mainstream LVLMs. Fig.6 illustrates the out-of-distribution performance of M-DRUM on DGM^4 dataset. It can be observed that PandaGPT frequently misjudges the authenticity or is unable to produce coherent words. In the absence of particular analysis, GPT-4 tends to propose ambiguous conclusions. Despite the in-depth explanation of the news content provided, LLaVA's analytical reasoning is not very accurate. While other models show unsatisfactory performance on manipulation reasoning, our model demonstrates proficiency in proposing reasoning combining the description of news content and the analysis on the potential manipulations. Furthermore, our model can generate analytical reasoning with more confidence and vividness, expanding the implementation of fake news detection models.

5 DETAILS OF HUMAN EVALUATION

In human evaluation, each volunteer is paid to rate the results of different models' reasoning on five randomly selected items. All of our 12 raters are professional and diverse in background. The process of evaluation is independent with samples randomly chosen and shuffled for evaluation to reduce the impact of subjectivity. Specific human rating of different models on manipulation reasoning is exhibited in the supplementary Excel tables due to the space limitation.



(a) accuracy (b) precision (c) recall (d) f1-score

Figure 1: Performance floating of few-shot learning.

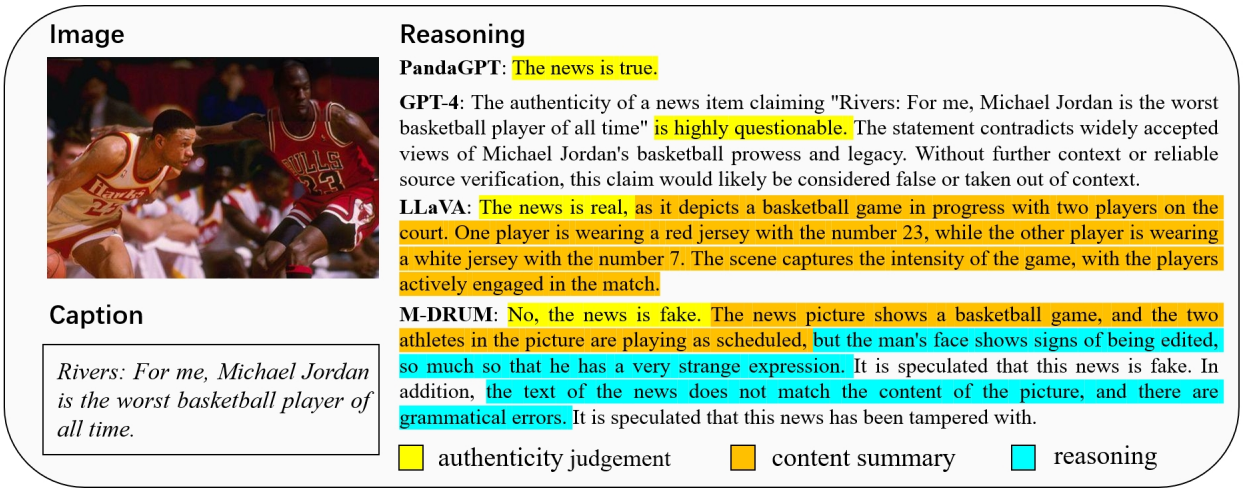


Figure 2: An example of manipulation reasoning in *sport* domain, which is visually and textually manipulated.

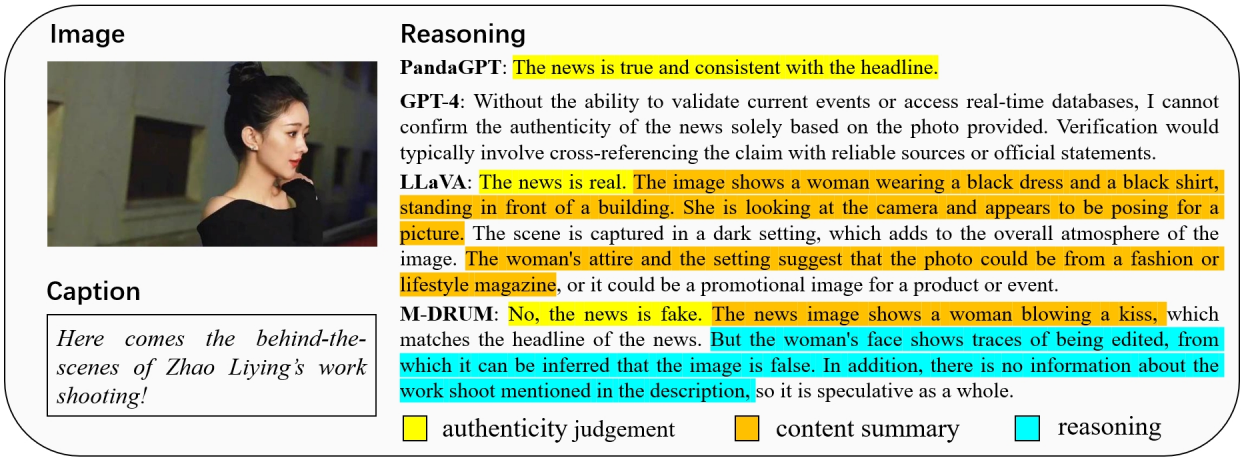


Figure 3: An example of manipulation reasoning in *entertainment* domain, which is visually manipulated

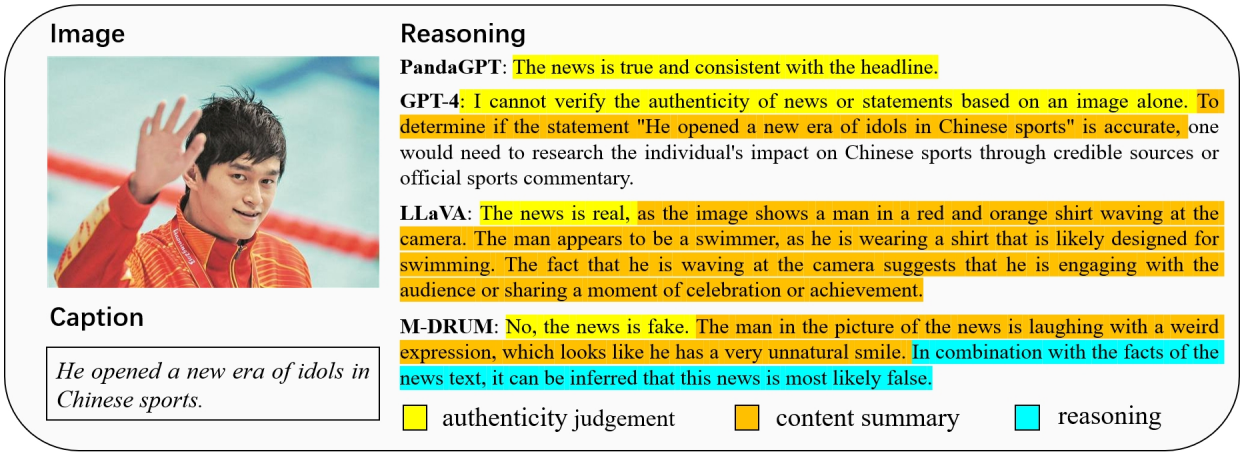


Figure 4: An example of manipulation reasoning in *sport* domain, which is visually manipulated.

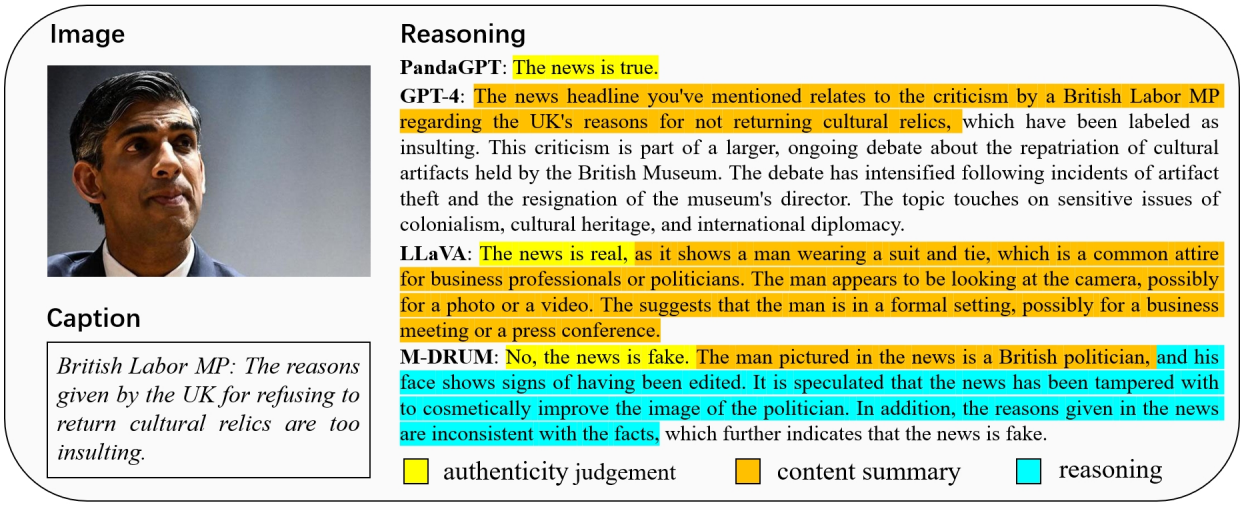


Figure 5: An example of manipulation reasoning in *politics* domain, which is visually and factually manipulated.

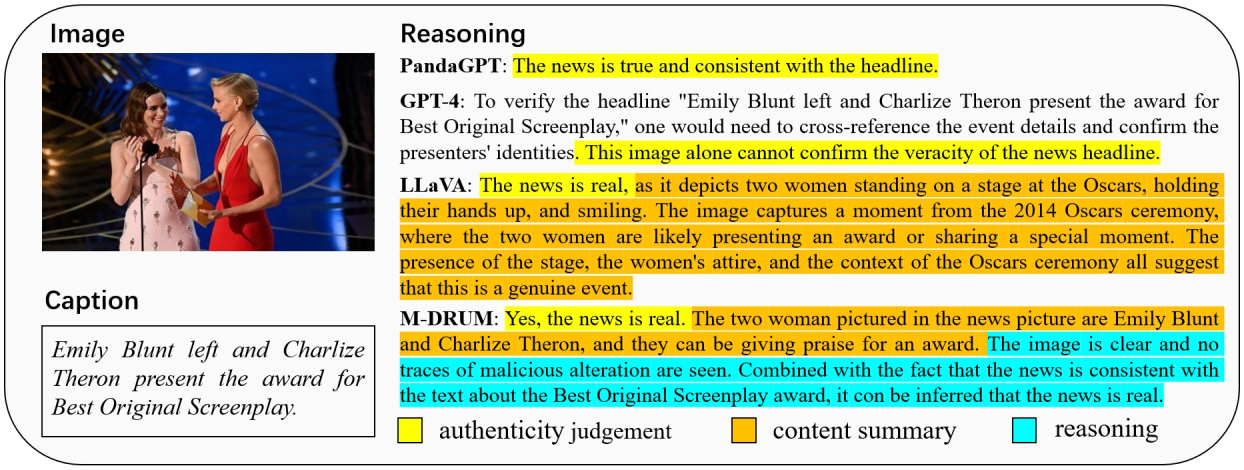


Figure 6: An Out-Of-Distribution example of manipulation reasoning in DGM^4 dataset, which is real.